IMPACT: International Journal of Research in Humanities, Arts and Literature (IMPACT: IJRHAL)
ISSN (P): 2347-4564; ISSN (E): 2321-8878
Vol. 4, Issue 6, Jun 2016, 25-32

ITEM ANALYSIS OF UNIVERSITY-WIDE MULTIPLE CHOICE OBJECTIVE EXAMINATIONS – THE EXPERIENCE OF A NIGERIAN PRIVATE UNIVERSITY

J.A. ODUKOYA, A.O. IGBINOBA, O. ADEKEYE & A. AFOLABI

Covenant University, Nigeria

ABSTRACT

What is not inspected should hardly be expected. It can be argued that it is regular inspection, constructive feedback and diligent correction that lead to growth and development. These statements give credence to the power of regular assessment and evaluation. Assessment is the umbrella term for all processes used to collect information from a person, thing or event for interpretative purposes. Tests and examinations are amongst the popular assessment tools. Item analysis is an integral tool for enhancing the content validity of multiple-choice objective examinations. Teachers and Students worldwide often dance to the tune of tests and examinations. Consequently, assessments are powerful tools for catalyzing the achievement of educational goals, especially if done rightly. It is the realization of the significance of rightly conducting assessment and evaluation to get the right results that the need to conduct the item analyses of the unitary multiple-choice objective semester examinations for university-wide courses conducted in a Nigerian private university was conducted. Consequently, the ex-post facto design was adopted for this project. Two core item analysis indices – item difficulty index and distractive index - were computed. Based on the findings from this study, particularly in the light of best practices in the art of psychometrics, recommendations were made to expedite the achievement of the 'laudable vision' of the investigated private university – *to be one of the best ten universities in the world by 2022*.

KEYWORDS: Item Analysis; Multiple Choice Objective Tests; Examination; Undergraduate; Nigeria; Private University.

INTRODUCTION

Background

Multiple choice objective tests items are easy to score and analyze but often technical, time consuming and at times painstaking in development. To cover a wide scheme of work or syllabus adequately, it is imperative that multiple-choice objective test be used. When assessing a large population of students, the use of multiple-choice question [MCQ] is the most logical option. The challenges however are: tendency to write poor MCQs with ambiguous prompts, poor distractors, multiple answers when question demands only one correct answer, controversial answers, give-away keys, higher probability of testees guessing correctly to mention but few of the challenges of developing and using MCQs.

There is hardly any subject that cannot use MCQ. However, when assessments border on life sensitive issues like health, air flight [and the like], it should be applied with caution. The reality, however, is that virtually all assessment purposes are life sensitive. The results of virtually all assessments are often used to make sensitive decision that determine people's destiny. It is therefore imperative that MCQs be handled rightly at the development, administration, scoring, grading and interpretation stages. The focus of the study reported here is on the development stage of MCQs, with

particular emphasis on item analyses.

The first critical step in developing valid MCQs is recruiting relevant subject experts with requisite skill in writing of MCQ items. The correct handling of this stage will go a long way in setting the pace for the establishment of the content validity of the test. However, the validity of MCQs cannot be completely ascertained with skillful item writing alone. Psychometric requirement demands that such items be trial tested, while the responses and scores generated are subjected to statistical item analyses.

There are three popular forms of item analyses: ***item difficulty index, distractive index*** and ***discriminatory index***. "Item analysis involves use of statistics that can provide relevant information for improving the quality and accuracy of multiple choice question". (Ary, Jacobs & Razavieh 2002; Boopathiraj & Chellamani, 2013; Boyle & Radocy, 1987; El-Uri & Malas, 2013; IAR, 2011; Sabri, 2013)

Item difficulty index indicates the degree of difficulty of the MCQ in relation to the cognitive ability of the testees. It is calculated by finding the proportion of the testees that got the item correctly. An item is adjudged too difficult when the index is below 0.3. An item is adjudged too easy when the index is above 0.7. Depending on the purpose of the test, the cut off points for easy or difficult items can be adjusted upward or downward. Generally, the rule is that life sensitive or competitive activities require more difficult items in screening; while less sensitive activities or activities requiring motivation of testees often use less difficult items. For most summative assessments, such as those handled by the West African Examinations Council, moderate difficulty index ranging around 0.5 are often preferred.

It is important to note that an item may record high difficulty index if the content of such item was not taught, the concept was not understood or if the question was not properly worded. This is actually the essence of item analysis – to check for flaws of this nature and find ways of correcting them before finally administering the questions. Item moderation, therefore, naturally follows item analysis. Where an item cannot be moderated, it is often discarded and replaced.

The distractive index determines the power of the distractor [i.e. the incorrect options in a MCQ] in distracting the testees. The distractive index is computed in virtually the same way as the difficulty index. It is the proportion of tesstes who selected a distractor out of all the testees that sat for the test. When a distractor distracts few or no testee, it is concluded that such is a poor distractor and should be reviewed. When a distractor over-distracts, that is, distracts about the same proportion or higher proportion of the testees that are selecting the key [i.e. right option], such option is also due for review or replacement.

Discriminatory index depicts the power of an item in discriminating between high and low performing Testees. Item discrimination determines whether those who did well on the entire test did well on a particular item. An item should in fact be able to discriminate between upper and lower scoring groups. One way to determine an item's power to discriminate is to compare those who have done very well with those who have done very poorly, known as the extreme group method. First, identify the Testees who scored in the top one-quarter [upper quartile] as well as those in the bottom one-quarter of the class [lowest quartile]. Next, calculate the proportion in the upper and lower quartiles that answered a particular test item correctly. Finally, subtract the proportion of Testees who got the item right in the bottom performing group from the proportion of Testees in the top performing group who got the item right to obtain the item's discrimination

index (D). Item discriminations of D =.50 or higher are considered excellent. D = 0 means the item has no discriminatory power, while D = 1.00 means the item has perfect discrimination power. It is therefore expected that more of the high performing Testees should get an item right while few of the low performing students should get the same item right. When more Testees who generally perform poorly in a test tend to select the right option for an item and those who performed well are selecting wrong options as answer, then something is apparently wrong with such an item. It calls for item review or discard. Thus, item analyses activities work to enhance the overall validity of a test.

Kehoe (1995) observed that the basic idea that we can capitalize on is that the statistical behavior of "bad" items is fundamentally different from that of "good" items. This fact underscores the point of view that tests can be improved by maintaining and developing a pool of "good" items from which future tests can be drawn in part or in whole. This is particularly true for instructors who teach the same course more than once. Item analysis is a tool to help the item writer improve an item (Gochyyev & Sabers, 2010)

Over the years, tertiary institutions have come to realize the significance of some life-enhancing concepts that should be learnt. It is these vital life-enhancing information that have been packaged as university wide courses. Consequently, some universities have compulsory courses like General Studies, which covers use of languages and philosophical issues; Total Man Concept; Entrepreneurship Development Studies; Human Development etc. Some of these courses are zero unit compulsory courses. The truth is that knowledge, especially applicable and relevant knowledge, are powerful and life transforming. It is therefore imperative to teach and assess these courses professionally for maximum impact. It is against the backdrop of these points this study was undertaken.

Statement of Problem

Inadvertent omission of item analysis in the course of developing Multiple Choice Questions [MCQ] for sensitive university wide courses that predominantly use MCQ could be a threat to the destiny of the Testees. Incorrect application of item analysis could yield the same fate. There were cases of students having to spend extra one year on the ground of failing some of these university-wide courses assessed with MCQs. This experience has implication for the fulfillment of affected students' destiny. The emotional offshoot of failing and having to spend an extra year with one's juniors could translate to a number of debilitating medical, psychosomatic and psychological challenges. On the other hand, unprofessional assessment could lead to wrong award of grades and certificates. The consequence of this practice is that graduates from such institutions may be incapable of favorably competing with students of same level from other institutions worldwide. These are problems that call for urgent attention, hence this study.

Statement of Significance

Professional conduct of item analysis and concomitant item moderation of items comprising the university wide courses is apt to enhance the overall validity of such tests. This in turn is apt to significantly reduce frustrations for the individual and the society at large. Correct assessment, with application of essential psychometric practices like item analysis is apt to enhance the quality of graduates from our institutions.

Statement of Objectives **[Purpose]**

- Find out how appropriate the difficulty indices of the items comprising the university wide courses are?
- Determine the appropriateness of the distractive indices of the options making up the items in the university-wide course MCQs?

Research Questions

- How appropriate are the difficulty indices of the items comprising the university wide courses?
- How appropriate are the distractive indices of the options making up the items in the university wide course MCQs?

METHOD

The research design adopted for this study is the ex-post facto design. Existing data were collated and analyzed.

The population for this study were undergraduates of private universities in Nigeria. They were estimated at about 3,000,000 as at the time of this study.

The responses of over 1500 students that responded to the MCQs of the university wide course at various times were harvested and analyzed. Students responses in following courses were analyzed: EXX 121 [N = 1907; Test taken 2015]; GXX 121; N = 1956; Test taken 2015]; HXX 421 [N = 112; Test taken 2015]; TXX 121 [N = 1905; Test taken 2015]. Note that original course codes have been changed for anonymity.

The core instruments for this study were the past MCQ items for four core university-wide courses.

Secondary data was collected from the computer department of the private university.

The major statistical analyses conducted were difficulty index and distractor index.

RESULTS & DISCUSSION

The following decision rules were applied to determine items that are Okay [OK], Fairly Okay [F/OK], Need Moderation [NM], and Need Serious Moderation [NSM]: When the difficulty index is over 0.7 [i.e.70%] or below 0.2 [i.e. 20%], such item is adjudged not okay and needs moderation. The difficulty index is computed with the proportion of Testees selecting the correct option as indicated by the bold figures in Table 1 below. When the distractive index for a distractor or incorrect option is far above or far below 0.166 [i.e. 16.6 %], there is also need for moderation. The rationale for this decision is that for a test that operates by the principle of moderate difficulty of 0.5, the remaining 0.5 should be fairly shared equally between the 3 distractors (for a 4-option item), which gives 0.166. Any item falling short of these two requirements is apt to require moderation.

Table 1: 2015 EXX 121 N = 1907

Items	A (%)	B (%)	C (%)	D (%)	E (%)	Comment
1	1.2	.5	5.1	16.6	**76.5**	Need Serious Moderation [NSM]
2	16.3	5.3	20.8	**35.7**	21.4	Need Moderation [NM]
3	**72.8**	1.9	6.7	17.6	.5	NSM
4	**90.0**	9.4	0.1	0.1	0.1	NSM

5	**37.3**	60.8	0.1	0.7	0.2	NSM
6	6.4	9.6	**17.2**	53.3	12.7	NSM
7	**15.7**	82.7	0.3	0.7	0.1	NSM
8	**30.3**	8.4	4.2	34.9	21.6	NSM
9	**80.0**	18.8	0.3	0.4		NSM
10	64.3	**26.4**	2.7	5.2	.2	NSM
11	**94.6**	3.0	2.2	0.1	0.1	NSM
12	**94.3**	4.7	0.4	0.1		NSM
13	4.6	**16.9**	35.1	42.6	0.2	NSM
14	**22.5**	1.6	2.8	1.9	71.1	NSM
15	**83.6**	15.5	0.1	0.2	0.1	NSM
16	**79.1**	19.9	0.2	0.2	0.2	NSM
17	1.5	1.4	9.0	41.6	**46.1**	NM
18	0.4	0.8	21.8	53.8	**22.9**	NSM
19	19.0	4.3	6.1	**62.7**	7.0	NM
20	11.6	48.9	13.6	**24.0**	0.6	NSM
21	20.5	14.7	**47.2**	16.7	0.4	OK
22	6.4	**20.5**	3.3	1.2	67.6	NSM
23	73.6	1.0	**24.5**	0.5	0.2	NSM
24	0.5	12.8	3.3	**81.5**	1.5	NM
25	**82.1**	16.3	0.3	0.3		NM
26	11.7	**49.7**	20.1	17.0		OK
27	**94.1**	5.2	0.2	0.1		NSM
28	**11.5**	19.1	18.2	47.4	3.1	NSM
29	**92.3**	6.6	0.2	0.2		NSM
30	40.5	**58.4**	0.2	0.4	0.2	F/OK
31	**67.1**	2.0	2.3	9.1	18.9	NM
32	1.7	6.2	43.1	21.2	**27.1**	NSM
33	**86.6**	12.6	0.3	0.1	0.2	NSM
34	12.0	11.1	13.6	3.3	**59.1**	F/OK
35	8.0	12.7	**64.9**	13.7	0.4	F/OK
36	23.7	23.6	23.0	**28.8**	0.1	NSM
37	**99.0**	0.8	0.1	0.1		NSM
38	42.7	**56.7**	0.1	0.2		NM
39	**88.4**	11.2	0.2	0.1	0.1	NSM
40	**92.6**	6.7	0.2	0.2		NSM
41	9.4	13.7	**75.4**	0.8		NM
42	**78.8**	15.6	3.5	1.7	0.1	NSM
43	14.1	**80.1**	1.0	4.3	0.1	NSM
44	**50.1**	13.6	3.6	32.2	0.1	OK
45	14.3	22.4	**41.6**	7.7	13.2	OK
46	12.5	22.5	25.6	19.9	**18.5**	NSM
47	**72.1**	3.8	5.9	10.5	7.3	NM
48	9.2	5.6	**10.4**	15.9	58.5	NSM
49	5.5	21.6	5.3	4.4	62.5	F/OK
50	31.0	38.5	27.7	2.3	**0.1**	NSM
51	**92.8**	1.6	0.6	4.7		NSM
52	3.0	94.2	1.1	**1.5**		NSM
53	39.7	15.6	26.1	**17.1**	0.1	NSM
54	**20.5**	51.7	15.8	10.5	0.1	NSM
55	**80.9**	3.7	10.0	4.7		NM
56	**42.2**	26.0	23.8	6.9	0.2	F/OK
57	26.7	58.0	**3.8**	10.5	0.1	NSM
58	12.5	**1.6**	1.5	83.5	0.1	NSM
59	**1.8**	3.0	88.9	5.4	0.2	NSM

60	70.1	**11.5**	6.6	10.3	0.3	NSM
61	23.3	**49.6**	23.6	3.1		F/OK
62	**0.8**	10.0	5.0	84.0		NSM
63	**6.0**	0.4	1.2	92.2	0.2	NSM
64	72.8	6.0	6.0	**10.4**	0.2	NSM
65	**7.7**	79.8	4.0	2.4	0.2	NSM
66	30.7	7.9	4.5	35.2	**21.1**	NSM
67	14.4	**29.7**	8.1	11.6	35.6	NSM
68	1.9	**9.1**	30.8	56.9	0.5	NSM
69	11.9	**5.9**	56.2	24.3	0.3	NSM
70	3.7	1.4	3.8	56.3	**34.6**	NM

Table 2: Summary for 2015 EXX 121 [1907 Students]

Description	Frequency [N=70]	%
1. Items that are **okay**	4	5.7%
2. Items that are **fairly okay**	6	8.6%
3. Items that **need moderation [NM]**	11	15.7%
4. Items that **need serious moderation [NSM]**	49	70%

The item analysis results here show that a significant majority of the items [approximately 86% of the 70 items fielded] did not meet psychometric standard [of appropriate difficulty and distractive index] and consequently need moderation, For a course that is deemed important enough to make it compulsory for the entire student body in a university, it imperative that it should be properly assessed. These results therefore call for attention.

Table 3: Summary for 2015 GXX 121 [1956 Students]

Description	Frequency [N=70 items]	%
1. Items that are **okay**	8	11.4%
2. Items that are **fairly okay**	16	22.9%
3. Items that **need moderation [NM]**	12	17.1%
4. Items that **need serious moderation [NSM]**	34	48.6%

The item analysis results here show that a notable proportion of the items [approximately 66% of the 70 items fielded] did not meet psychometric standard [of appropriate difficulty and distractive index] and consequently need moderation, The result appears better than the EXX 121 result above. However, for an important university-wide compulsory course of this nature, there is need to review the assessment procedure for enhancement of psychometric standards.

Table 4: Summary for 2015 HXX 421 [112 Students]

Description of Items	Frequency [N=70 items]	%
1. Items that are **okay**	--	
2. Items that are **fairly okay**	2	2.9%
3. Items that **need moderation[NM]**	15	21.4%
4. Items that **need serious moderation [NSM]**	53	75.7%

The item analysis results here show that a notable proportion of the items [approximately 97.1% of the 70 items fielded] did not meet psychometric standard [of appropriate difficulty and distractive index] and consequently need moderation. Only 2.9% of the items were fairly okay. The operational psychometric standard for this study is that at least 70% of the items should be *okay* while the remaining 30% could be *fairly okay*. It is unacceptable to administer items having poor distractors and give-away keys as implied with item analysis results of keys featuring above 70% and distractors featuring below 16% selection for a 4 options MCQ.

For an important university-wide compulsory course of this nature, there is need to review the assessment procedure for enhancement of psychometric standards.

Table 5: Summary for 2015 TXX 121 [1905 Students]

Description of Items	Frequency [N=70 items]	%
1. Items that are **okay**	2	2.9%
2. Items that are **fairly okay**	10	14.3%
3. Items that **need moderation [NM]**	11	15.7%
4. Items that **need serious moderation [NSM]**	47	67.1%

The item analysis results here show that a significant majority of the items [approximately 83% of the 70 items fielded] did not meet psychometric standard [of appropriate difficulty and distractive index] and consequently need moderation. For a course, deemed important enough to make it compulsory for the entire student body in a university, it demands that better psychometric standards should be attained and maintained. These results seriously call for attention.

RECOMMENDATIONS AND CONCLUSION

On the strength of the findings made from this study, it recommended that the development of all the university-wide courses employing the MCQ format should commence with preparation of test blueprint, items should be carefully written following the rules for writing multiple-choice objective questions [MCQs], all items should be trial tested, item analyzed and subjected to item moderation to enhance the overall *content* and *construct* validities. This exercise will require the input of subject and psychometric experts. The exercise should be statutory for quality assurance. Dogged adoption of this singular recommendation is apt to significantly enhance quality of certification and ultimately the quality of graduates from the respective tertiary institutions.

REFERENCES

1. Ary, D., Jacobs L.C. & Razavieh, A. (2002). *Introduction to Research in Education*. (6th ed.). California: Wadsworth.
2. Boopathiraj, C. & Chellamani, K. (2013). Analysis of Test Items on Difficulty Level and Discrimination Index in the Test for Research in Education. *International Journal of Social Science and Interdisciplinary Research*; Vol 2:2
3. Boyle, J.D & Radocy, R.E. (1987). *Measurement and Evaluation of Musical Experiences*. NewYork: Macmillan.

4. El-Uri, F.I & Malas, N. (2013). Analysis of Use of A Single Best Answer Format in An Undergraduate Medical Examination. *Qatar Medical Journal* 2013:1.

5. Gochyyev, Perman & Sabers, Darrell (2010). Item Analysis in *Journal of Research Methods.* SAGE Online Publications: https://srmo.sagepub.com/view/encyc-of-research-design/n199.xml

6. Instructional Assessment Resources (2011). Item Analysis. Retrieved November 9, 2013 from University of Texas at Austin, *Instructional Assessment Resources*, IAR Web site: http://www.utexas.edu/academic/ctl/assessment/iar/students/report/itemanalysis.php

7. Kehoe, Jerard (1995). Basic item analysis for multiple-choice tests. *Practical Assessment, Research & Evaluation*, 4(10). Retrieved September 26, 2015 from http://PAREonline.net/getvn.asp?v=4&n=10

8. Sabri, Shafizan (2013). 'Item Analysis Of Student Comprehensive Test For Research In Teaching Beginner String Ensemble Using Model Based Teaching Among Music Students In Public Universities' In *International Journal of Education and Research* Vol. 1 No. 12

9. Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum